A distributional study of negated adjectives and antonyms

Aina, L.; Bernardi, R.; Fernández, R.

Publication date
2018

Document Version
Final published version

Published in
CLiC-it 2018 : Italian Conference on Computational Linguistics

Citation for published version (APA):
A distributional study of negated adjectives and antonyms

Laura Aina
Universitat Pompeu Fabra
Barcelona, Spain
laura.aina@upf.edu

Raffaella Bernardi
University of Trento
Trento, Italy
bernardi@disi.unitn.it

Raquel Fernández
University of Amsterdam
Amsterdam, The Netherlands
raquel.fernandez@uva.nl

Abstract

English. In this paper, we investigate the relation between negated adjectives and antonyms in English using Distributional Semantics methods. Results show that, on the basis of contexts of use, a negated adjective (e.g., not cold) is typically more similar to the adjective itself (cold) than to its antonym (hot); such effect is less strong for antonyms derived by affixation (e.g., happy - unhappy).

Italiano. In questo lavoro, analizziamo la relazione fra aggettivi negati e antonimi in inglese utilizzando metodi di Semantica Distribuzionale. I risultati mostrano che, sulla base dei contesti di uso, la negazione di un aggettivo (ad es. “not cold”; it.: “non freddo”) è tipicamente più simile all’aggettivo stesso (“cold”; it.: “freddo”) che al suo antonimo (“hot”; it.: “caldo”). Tale effetto è meno accentuato per antonimi derivati tramite affissi (ad es. “happy”-“unhappy”; it.: “felice”-“infelice”).

1 Introduction

Negation has long represented a challenge for theoretical and computational linguists (see Horn (1989) and Morante and Sporleder (2012) for overviews): in spite of the relative simplicity of logical negation (¬p is true ↔ p is false), complexity arises when negation interacts with morphosyntax, semantics and pragmatics.

In this work, we focus on the negation of adjectives in English, expressed by the particle not modifying an adjective, as in not cold. A naive account of these expressions would be to equate them to antonyms, and hence take them to convey the opposite of the adjective (e.g., not cold = hot). In fact, this simplifying assumption is sometimes made in computational approaches which model negation as a mapping from an adjective to its antonym (e.g., The Pham et al., (2015), Rimell et al., (2017)). However, a range of studies support what is known as mitigation hypothesis (Jespersen, 1965; Horn, 1972; Giora, 2006), according to which a negated adjective conveys an intermediate meaning between the adjective and its antonym (e.g., not large = medium-sized). The meaning of the adjective is mitigated by negation, while some emphasis on it still persists in memory (Giora et al., 2005). This view is compatible with pragmatic theories predicting that the use of a more complex expression (not large) when a simpler one is available (small) triggers the implicature that a different meaning is intended (e.g., medium-sized) (Grice, 1975; Horn, 1984). Computational models predicting similar mitigating effects are those by Hermann et al., (2013) and Socher et al., (2012; 2013).

In this work, we investigate negated adjectives from the perspective of Distributional Semantics (Lenci, 2008) Turney and Pantel, 2010). We study antonymic adjectives and their negations in terms of their distribution across contexts of use: to this end, we employ an existing dataset of antonyms, whose annotation we further extend, and the distributional representations of these and their negated version, as derived with a standard distributional model. This allows us to conduct a data-driven study of negation and antonymy that covers a large set of instances. We compare pairs of antonyms with distinct lexical roots and those derived by affixation, i.e., lexical and morphological antonyms (Joshi, 2012) (e.g., small - large and happy - unhappy respectively). More-

* Part of the work presented in this paper was carried out while the first author was at the University of Amsterdam.
over, we investigate the distinction between lexical antonyms that are contrary or contradictory, that is, those that have or do not have an available intermediate value (Fraenkel and Shul, 2008): e.g., something not cold is not necessarily hot - it could be lukewarm - but something not present is absent. As for negations of morphological antonyms, we compare instances of simple and double negation, where the latter occurs if the antonym that is negated is an affixal negation (e.g., not unhappy).

Our analyses show that, when considering distributional information, negated adjectives are more similar to the adjective itself than to the antonym (e.g., not cold is closer to cold than to hot), regardless of the type of antonym or of negation. However, we find that morphological antonymy is closer to negation than lexical one is.

2 Motivation and data

We are interested in how negation acts with respect to pairs of adjectives connected by the lexical relation of antonymy (Murphy, 2003), i.e., that are associated with opposite properties within the same domain (e.g., hot - cold). In particular, we want to compare the negation of one of the antonymic adjectives with itself and its antonym respectively (e.g., not cold vs. cold and vs. hot). Our data of interest are then triples obtained starting from an antonymic pair and negating one of the two items (for each pair we obtain two triples). For example:

1. \( \langle \text{hot}, \text{cold}, \text{not} \{\text{hot}|\text{cold}\} \rangle \)
2. \( \langle \text{happy}, \text{unhappy}, \text{not} \{\text{happy}|\text{unhappy}\} \rangle \)

As data, we make use of a subset of the Lexical Negation Dictionary by Van Son et al. (2016). This consists of antonym pairs in WordNet (Fellbaum, 1998) annotated for different types of lexical negation (Joshi, 2012). We consider adjective pairs that are either lexical antonyms, i.e., with distinct lexical roots (e.g., cold - hot), or morphological antonyms, i.e., derived by affixal negation (e.g., happy - unhappy)\(^1\). In our analyses, we compare different subsets of the data: we explicate and motivate the distinctions in the following.

Lexical vs. morphological antonyms These two groups are usually taken to express the same lexical relation - i.e., opposition - and to be different only on morphological terms. However, such

<table>
<thead>
<tr>
<th>Type</th>
<th>adj.</th>
<th>not_adj.</th>
<th># triples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical antonyms</td>
<td>254715</td>
<td>1144</td>
<td>198</td>
</tr>
<tr>
<td>– contrary</td>
<td>336923</td>
<td>1057</td>
<td>68</td>
</tr>
<tr>
<td>– contradictory</td>
<td>298378</td>
<td>1031</td>
<td>28</td>
</tr>
<tr>
<td>Morphological antonyms</td>
<td>83232</td>
<td>1821</td>
<td>185</td>
</tr>
<tr>
<td>– simple negations</td>
<td>84744</td>
<td>2002</td>
<td>157</td>
</tr>
<tr>
<td>– double negations</td>
<td>122525</td>
<td>871</td>
<td>28</td>
</tr>
</tbody>
</table>

Table 1: Average frequency of adjectives and negated adjectives per class, and total number of triples \( \langle a_1, a_2, \text{not} \{a_1|a_2\} \rangle \) considered.

difference might affect their relation with negated adjectives: indeed, affixal negations have a morphological structure that resembles negated adjectives (e.g., unhappy vs. not happy). For this reason, we keep triples derived from lexical and morphological antonyms distinct, and compare them in our analyses: in particular, we are interested in testing whether in a distributional space negation tends to be more similar to morphological antonymy than to lexical one. Besides this comparison, we apply other distinctions to the triples obtained with lexical and morphological antonyms respectively, in order to investigate further effects.

Contrary vs. contradictory Lexical antonyms have been classified as either contrary or contradictory (Clark, 1974), depending on whether the negation of one entails the truth of the other, without the availability of a mid-value. Fraenkel and Shul (2008) provided psycholinguistic results showing that if an adjective is part of a contradictory pair, its negation is interpreted as closer to the antonym than if it is part of a contrary pair (e.g., not dead is interpreted as being closer to alive than not small to large). We aim to investigate this result in a distributional space, where we are able to quantify similarities between lexical items.

Since no data annotated with respect to this distinction is available, the three authors independently annotated the antonym pairs in the dataset as either contrary, contradictory or unclear, following the definition used by Fraenkel and Shul (2008).\(^2\) Not surprisingly, the inter-annotator agreement is only moderate (Fleiss’ \(k = 0.37\)): already Fraenkel and Shul (2008) noted that even for what they considered contradictory pairs it is possible to conceive a mid-value interpretation (e.g., not dead \(\approx\) half-dead; Paradis and Willners (2006)). This suggests that the contrary

\(^1\)In the dataset, the former are coded as regular antonyms and the latter as direct affixal negations.

\(^2\)Annotation guidelines at https://lauraina.github.io/data/notadj.pdf
vs. contradictory distinction involves a continuum rather than a dichotomy. We leave this aspect to be further clarified by future research and, for the purpose of our analysis, only consider pairs classified with full agreement.

**Simple vs. double negation** In the case of morphological antonyms, one of the two adjectives is an affixal negation, and hence already contains a negating prefix (such as un- in unhappy): adding not thus gives rise to a double negation (e.g., not unhappy). These expressions have been widely studied in the literature due to their difference with double negation in logic (e.g., Bolinger (1972), Krifka (2007) and recently Tessler and Franke (2018)). While in logic two negations cancel each other out (\(\neg\neg p = p\)), in natural language double negations are typically employed to weaken the meaning of the adjective that is negated twice (e.g., not unhappy \(\neq\) happy). Our goal is to test whether evidence for this effect is found in a distributional space: in particular, if two negations were to cancel each other out then the negation of an affixal negation (e.g., not unhappy) should be particularly close to the antonym (e.g., happy). We then test whether simple (e.g., not happy) and double (e.g., not unhappy) negations exhibit similar trends in relation to an antonym pair (happy vs. unhappy).

### 3 Analyses

#### 3.1 Methods

Previous studies about negation of adjectives described its effect as a meaning shift from the adjective towards the antonym, that can be measured in terms of semantic similarity (Fraenkel and Schul. 2008). Distributional Semantics offers us a data-driven method of quantifying this: we can represent expressions as vectors summarizing their large-scale patterns of usage and then interpret their proximity relations in terms of similarity.

To this aim, we build a distributional semantic model with standard techniques, but whose vocabulary includes, besides word units, also negated adjectives. In practice, each occurrence of a negated adjective (adjacent occurrence of not and an adjective without intervening words; e.g., we exclude cases like not very cold) is treated as a single and independent token (e.g., not cold \(\sim\) not_cold). With this pre-processing, we train a word2vec CBOW model (Mikolov et al., 2013) on the concatenation of UkWaC and Wackypedia-En corpora (2.7B tokens; Baroni et al., (2009)), setting parameters as in the best performing model by Baroni et al. (2014). We do not carry out any hyperparameters search, nor we employ any ad hoc techniques aimed at, for example, amplifying the distances between antonyms in the semantic space (such as that of Nguyen et al. (2016) or The Pham et al. (2015)). Indeed, we are interested in investigating characteristics of antonyms and negated adjectives in a standard distributional model, that is not fine-tuned to a particular task and where no assumptions about the structure of its space are incorporated. However, we assess the quality of the induced model through a similarity relatedness task, where we find that it achieves satisfying performances.

For our analyses, we consider triples as those described in Section 2. Given a triple \((a_i, a_j, \text{not} a_i)\) (e.g., cold, hot, not cold), we define the following score:

\[
\text{Shift} := \text{Sim}(\text{not} a_i, a_j) - \text{Sim}(\text{not} a_i, a_i)
\]

where \(i \neq j\), and \(\text{Sim}(\text{not} a_i, a_j)\) and \(\text{Sim}(\text{not} a_i, a_i)\) are the cosine similarities of the negated adjective with the antonym and the adjective, respectively. This measures how much closer a negated adjective is to the antonym than to the adjective (i.e., how much closer not cold is to hot than to cold), and hence how much negation shifts the meaning of an adjective towards that of the antonym. Due to the well-known tendency of antonyms to be close in a distributional space (Mohammad et al., 2013), the absolute value of Shift is not expected to be high (a vector close to one is likely close to the other too). However, we can test whether a higher proximity is registered towards one of the two adjectives.

From the data introduced in Section 2, we only consider triples where each of the three elements occurs at least 100 times in the training corpus of the distributional model. Table 1 shows the number of triples considered for each class and the average frequency of adjectives and negated adjectives. The number of contradictory triples is

\[
\text{Gensim implementation.}
\]

\[
\text{Vectors size: 400; window size: 5; minimum frequency: 20; sample: 0.005; negative samples: 1.}
\]

\[
\text{Spearman’s} \rho \text{ of 0.75 on the MEN dataset (Bruni et al., 2014); see results by Baroni et al. (2009) for a comparison.}
\]

\[
\text{Negated adjectives are overall less frequent than their non-negated counterparts, as shown in Table 1.}
\]
3.2 Results and discussion

Table 2 shows the scores across the different categories mentioned in Section 2. Example triples for each category are given in Table 3 together with the nearest adjectives of each element in the triple.

Lexical vs. morphological antonyms The average Shift scores of both classes are negative, showing that a negated adjective is typically closer to the adjective than to the antonym. Indeed, as shown in Table 3, the nearest neighbor of a negated adjective is often the related adjective. On one hand, this could be seen as supporting the idea that negated adjectives express an intermediate meaning between that of the adjective and the antonym (e.g., not small is close to normal-sized). More in general, it shows that negated adjectives have a profile of use that is more similar to that of the adjective than to the antonym.

The two classes of antonyms differ significantly in the extent of this effect: negated adjectives are closer to a morphological antonym than a lexical one (e.g., not perfect vs. imperfect, not wide vs. narrow). Such similarity in distribution can be explained by the similarity in structure, and hence possibly in meaning, of negated adjectives and affixal negations. Yet, in spite of the higher similarity in use, affixal negation still does not seem equivalent to negation by not, due to the negative average Shift value.

Contrary vs. contradictory antonyms In contrast to the results from the linguistic literature (see Section 2), the behavior of contrary and contradictory antonym pairs is not significantly different in our analysis. When we look into a distributional space, even for contradictory antonyms, the negated adjectives tend to be more similar to the adjective itself than to the antonym.

This result points to the fact that distributional similarity is capturing a different type of similarity from that considered in the experiments of Fraenkel and Shul (2008). We cannot thus directly interpret our results as just a product of the mitigating aspect of negation. Distributional information may discriminate between the negation of an adjective and the antonym, even when the two seem intuitively equivalent (e.g., not dead is closer to dead than to alive): indeed, the use of one or the other may serve different functions (e.g., contradicting an expectation, politeness, etc.), leading them to appear in different contexts. Moreover, we find that, since continuous representations are able to capture nuanced differences, the alleged dichotomy between contrary and contradictory antonyms may become a continuum in distributional space: for example, one of the closest adjectives to not dead is half-dead. This further underscores the difficulty in distinguishing between contrary and contradictory antonyms which we had already encountered in the annotation.

Simple vs. double negations There is not a significant difference between negated adjectives that are instances of simple and double negations: crucially, it is not the case that double negations are very close to the antonym as a result of the two negations canceling each other out (e.g., not unhappy is closer to unhappy than to happy).

As before, the result cannot be interpreted only in terms of mitigation (though, e.g., not unhappy is close to unimpressed, hence a mid-value between happy and unhappy). In general, it suggests that the contexts of use of double negations are more similar to the ones of the adjective that is negated than to those of its antonym. Indeed, double negations typically appear in contexts where the use of the “logically” equivalent alternative (i.e., the antonym) is to be avoided for pragmatic reasons, as possibly too strong or direct (e.g., not unproblematic vs. problematic; Horn, 1984).

4 Conclusion

We have investigated negated adjectives using the tools of Distributional Semantics, which allows us to quantify the similarities between expressions on the basis of how they are used. Our analyses show that, when considering contexts of occurrence, negating an adjective does not make it closer to the antonym than to the adjective itself. This can be seen as a result of the various functions of negation (e.g., mitigation, contradiction to an expectation, politeness) that may lead to different patterns of use for negated adjectives and antonyms. Further research may shed light on which type of contexts actually discriminate them, for example through a corpus study, and which other properties negated adjectives have in a distri-

7 Full list of triples at https://lauraina.github.io/data/notadj.pdf
Table 2: Average Shift scores, with standard deviation, for each category. ***: significant difference between categories in the row ($p < 0.001$, Welch’s $t$-test).

<table>
<thead>
<tr>
<th>Category</th>
<th>Shift Score</th>
<th>Standard Deviation</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical antonyms</td>
<td>$-0.19$ ($\sigma = 0.16$)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Morphological antonyms</td>
<td>$-0.04$ ($\sigma = 0.16$)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contrary antonyms</td>
<td>$-0.18$ ($\sigma = 0.15$)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contradictory antonyms</td>
<td>$-0.19$ ($\sigma = 0.16$)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simple negations</td>
<td>$-0.03$ ($\sigma = 0.17$)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Double negations</td>
<td>$-0.06$ ($\sigma = 0.11$)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Nearest adjectives is semantic space for the three elements in some sample triples.

<table>
<thead>
<tr>
<th>Contrary antonyms</th>
<th>Small: large, tiny, smallish, sizeable, largish</th>
<th>Large: small, sizeable, huge, vast, smallish</th>
<th>Not small: small, smallish, normal-sized, largish, middle-sized</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contradictory antonyms</td>
<td>Dead: drowned, lifeless, half-dead, wounded, alive</td>
<td>Alive: dead, awake, unharmed, beloved, tortured</td>
<td>Not dead: dead, half-dead, alive, comatose, lifeless</td>
</tr>
<tr>
<td>Simple negations</td>
<td>Similar: analogous, identical, comparable, dissimilar, same</td>
<td>Dissimilar: similar, different, distinct, unrelated, identical</td>
<td>Not similar: similar, dissimilar, identical, distinguishable, analogous</td>
</tr>
<tr>
<td>Double negations</td>
<td>Happy: glad, pleased, contented, nice, kind</td>
<td>Unhappy: disappointed, dissatisfied, unsatisfied, resentful, anxious</td>
<td>Not unhappy: unhappy, adamant, disappointed, dismayed, unimpressed</td>
</tr>
</tbody>
</table>

butional space, such as their interaction with scalar dimensions (e.g., not hot vs. freezing, cold, lukewarm, hot etc.; Wilkinson and Tim (2016)). Finally, while for the purpose of this study we opted for a standard word2vec model, one could test for the same effects with differently obtained distributional vectors.

Despite its current limitations in covering truth-related aspects of meaning, Distributional Semantics was shown by Kruszewski et al. (2017) to be apt to model at least some aspects of negation, especially if graded in nature, such as alternativehood. Our study provides supporting evidence for this line of research and in addition points at the utility of using Distributional Semantics to uncover nuanced differences in use between a negation and other expressions, even when logically equivalent. Moreover, we regard our results to be of general interest for the NLP community, since effects of negation like the ones we studied and how they are represented in a distributional space can be critical for tasks like sentiment analysis (e.g., what does it imply that a costumer is not happy or not unhappy with a product?; Wiegand et al. (2010)).

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No 715154), and by the Catalan government (SGR 2017 1575). This paper reflects the authors’ view only, and the EU is not responsible for any use that may be made of the information it contains.

References


Germán Kruszewski, Denis Paperno, Raffaela Bernardi, and Marco Baroni. 2017. There is no logical negation here, but there are alternatives: Modeling conversational negation with distributional semantics. *Computational Linguistics*.


